

本周周报（2014.04.14-2014.04.20）

郭方舟

本周工作

1 空气污染数据可视化

本周与斐然师兄讨论了接下来需要实现的东西：

1. 加入 CCA 分析
2. 传感器网络数据的分析
3. 可调节的 zoom in 和 zoom out

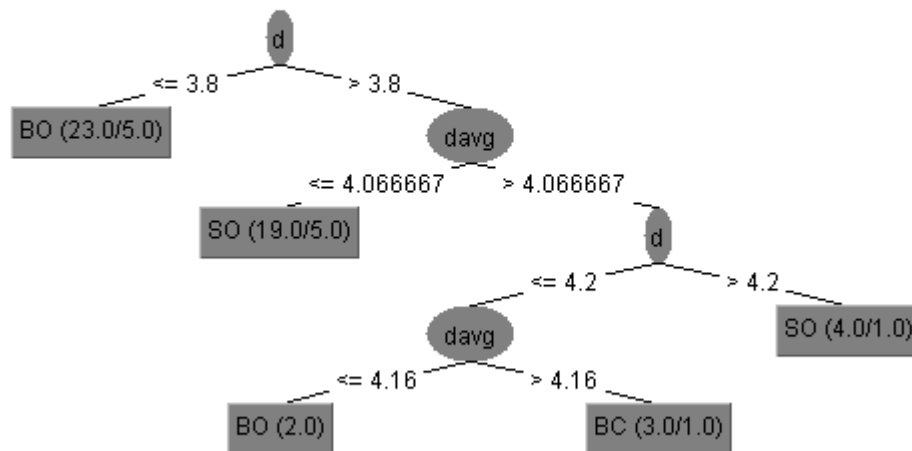
本周首先考虑的是加入 cca 分析。之前实现的程序实在太耗用内存，因此本周重新将数据放到了 mysql 数据库中，通过分页表和分开建立数据表的方法将查询速度提升到一秒二十万条数据左右。同时，要进行多属性 cca 分析，就要考虑量纲的问题，因此也去研究了一下去量纲的方法，其中均值化似乎是个不错的方法，但是还有待调研。

2 期货交易数据

目前，我们计算了 $d(10lastprice)$ （以 $d10$ 表示）， $d(11lastprice)$ （以 $d11$ 表示）， $(11lastprice - 10lastprice)$ （以 d 表示）， $d(11lastprice - 10lastprice)$ （以 dd 表示）和 $avg30(11lastprice - 10lastprice)$ （以 $davg$ 表示）五个值，并以给出的 15 分钟内的交易记录作为训练数据生成了决策树。对于交易类型，以 BO 代表买入开仓，SO 代表卖出开仓，BC 代表买入平仓，SC 代表卖出平仓。在生成决策树的过程中，暂时没有对成功和未成功交易记录进行区分。

以原始交易记录作为训练集生成的决策树如下图所示：

Tree View

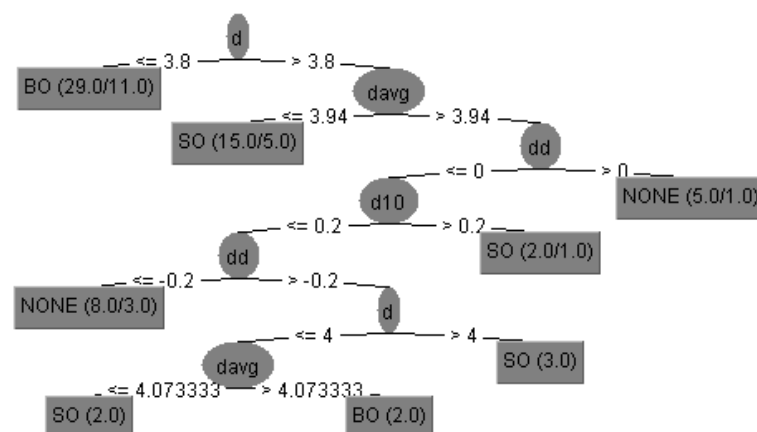


生成的决策树准确度为：

Correctly Classified Instances	33	64.7059 %
Incorrectly Classified Instances	18	35.2941 %

从决策树中可以清楚的看到行为似乎与差分量没有关系而只与两个差值有关。这种情况很可能是由训练集过小造成的，因此在 action 中增加第五个动作，**NONE**，代表不进行任何操作。随机选择 15 个无操作时间点置为 **NONE** 后，决策树变为：

Tree View



生成的决策树准确度为：

Correctly Classified Instances	33	50	66%
Incorrectly Classified Instances	33	50	33%

在人工添加了一些点之后，我们可以看到 $d(10lastprice)$ 和 $d(11lastprice - 10lastprice)$ 都被考虑进决策树，这个决策树的准确率为 50%。

下一步可以将决策树带入到没有交易记录的数据中进行模拟，观察能否盈利。

现在得到的决策树还不是非常准确，看起来也不太合理，这一方面是因为训练集太小，如果有更多的交易数据应该能得到更好的结果，另一方面是因为应该还有一些量没有纳入考虑，导致不准确。